Assignment 1

All teams have been assigned dataset.

You must do the following things with the dataset

1) ***Data Profiling***
2) ***Data Model Card***
3) ***Data as a Service***
4) ***Data Validation***

Note:

1) You need to create a GitHub Organization naming (***BigDataIA_Summer2022_ {Team Number}***) and create a repository inside it as ***Assignment1***

2) Make sure there are no errors in your python files when you submit it.

3) Have a requirement.txt file in your repository so we can install packages which you have used.

4) Whatever data you use needs to be staged somewhere(Cloud Platforms) and then you need to use it in your assignment files.

Data Profiling:

Use pandas profiling on your dataset

https://github.com/ydataai/pandas-profiling

Note: If you have an image dataset create a meta data of your dataset and do profiling on that.

For creating meta data, you can use:

https://pypi.org/project/Pillow/2.2.1/

https://www.thepythoncode.com/article/extracting-image-metadata-in-python

Published a py file which has an example of generating metadata for a dataset:

https://github.com/shahparth0007/Big-Data-Systems-Intelligence-Analytics-Labs-Summer-2022/tree/main/Generate_MetaData

Why Profiling: So, you know how your data looks and we are aware of all data outliers which we need to handle in great expectation

Submission: Create a folder in your GitHub repository where you have

1) ***IPYNP file containing the code of your pandas profiling***
2) ***HTML file containing the output of your profiling***

Data Model Card:

We will be using scikit-learn to create a model card.

Why Model Card? All the datasets have a purpose of solving some problem which can be a machine learning model. So, we want to tell our users some details about the same as for what we will be using in this dataset, what are user groups etc.

https://cloud.google.com/blog/products/ai-machine-learning/create-a-model-card-with-scikit-learn

Submission: Create a folder in your GitHub repository where you have

3) ***IPYNP file containing the code of your model card***
4) ***HTML file containing the output of your model card***

Data As a Service

In this Module you must create 5 – 10 Python functions which have input parameters and output as an image from your dataset

Requirements for each function:

1) Each python function should be in separate py file
2) Write the function Documentation: This should be written in plain English telling what the function will do with created date and author's name. Write it inside your py file

   https://realpython.com/documenting-python-code/

3) Write the Function Unit Test: Unit test for your python functions use pytest for this.

   https://docs.pytest.org/en/7.1.x/

4) Error Handling: Your functions should have try and catch statements to handle errors.

Submission: Create a folder in your GitHub repo and should have all the above files inside it

Data Validation:

In this module you will need to use the package Great Expectations, to create validations for your metadata of your image dataset.

Why Validation: When there is new data entry to your existing dataset you need to flag if there is any data which is not valid, so you do not process that.

Requirements:

1) Create at least 8 expectations for your dataset

Submission: Create a folder in your GitHub Repo and push the great expectations folder which you have created in this. It should contain the HTML report which shows what expectations you have created. (You can see the demo video for this).

You should also validate the expectations on your dataset.

Overall Requirements:

1) *Create a Claat Document which you will be presenting in class on 18th June 2022 which describes everything you did.*
2) *In your GitHub you should have a readme.md files which would tell what all things are there in this GitHub repository*
3) *Put an attestation in your readme.md file as below*

   *"**WE ATTEST THAT WE HAVEN'T USED ANY OTHER STUDENTS' WORK IN OUR ASSIGNMENT AND ABIDE BY THE POLICIES LISTED IN THE STUDENT HANDBOOK***

   ***Contribution: member1: 50% member2: 50%** "*

4) *You need to assign contribution % to your group depending on the work done by the person. This will be used to grade you and your team.*
5) *Please keep your repository private till the submission and make it public after the submission date and time.*
6) *Make sure you do not push anything to your GitHub after submission date (Editing Readme.md is ok but no code pushing after deadline)*